

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号
WO 2024/187991 A1

(43) 国际公布日
2024 年 9 月 19 日 (19.09.2024)

WIPO | PCT

(51) 国际专利分类号:
G06N 5/04 (2023.01)

(21) 国际申请号: PCT/CN2024/075820

(22) 国际申请日: 2024 年 2 月 4 日 (04.02.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310269966.4 2023年3月16日 (16.03.2023) CN

(71) 申请人: 杭州阿里云计算有限公司(HANGZHOU ALICLOUD FEITIAN INFORMATION TECHNOLOGY CO., LTD.) [CN/CN]; 中国浙江省杭州市余杭区五常街道文一西路969号3幢5层553室, Zhejiang 311121 (CN)。

(72) 发明人: 丁宇 (DING, Yu); 中国北京市朝阳区望京东园四区9号楼-阿里中心, 望京A座, Beijing

100102 (CN)。李鹏 (LI, Peng); 中国北京市朝阳区望京东园四区9号楼-阿里中心, 望京A座, Beijing 100102 (CN)。游亮 (YOU, Liang); 中国北京市朝阳区望京东园四区9号楼-阿里中心, 望京A座, Beijing 100102 (CN)。

(74) 代理人: 广州铸智知识产权代理有限公司 (GUANGZHOU ZHUZHI INTELLECTUAL PROPERTY LIMITED); 中国广东省广州市天河区林和西路161号中泰国际广场B座2305室, Guangdong 510620 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN,

(54) Title: INFERENCE METHOD AND SYSTEM FOR COMPUTATIONAL GRAPH, AND STORAGE MEDIUM

(54) 发明名称: 计算图的推理方法、系统及存储介质

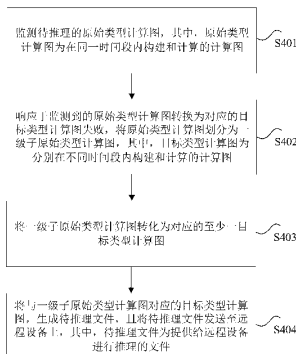


图 4

- S401 Detect an original-type computational graph to be inferred, wherein the original-type computational graph is a computational graph that is constructed and computed within the same time period
- S402 In response to a failure in the conversion of the detected original-type computational graph to a corresponding target-type computational graph, divide the original-type computational graph into primary original-type sub-computational-graphs, wherein the target-type computational graph is a computational graph that is constructed and computed within different time periods
- S403 Convert the primary original-type sub-computational-graphs into at least one corresponding target-type computational graph
- S404 On the basis of the target-type computational graph corresponding to the primary original-type sub-computational-graphs, generate a file to be inferred, and send said file to a remote device, wherein said file is a file that is provided for the remote device for inference

(57) Abstract: Disclosed in the present application are an inference method and system for a computational graph, and a storage medium. The method comprises: detecting an original-type computational graph to be inferred; in response to a failure in the conversion of the detected original-type computational graph to a corresponding target-type computational graph, dividing the original-type computational graph into primary original-type sub-computational-graphs; converting the primary original-type sub-computational-graphs into at least one corresponding target-type computational graph; on the basis of the target-type computational graph corresponding to the primary original-type sub-computational-graphs, generating a file to be inferred, and sending said file to a remote device. The present application solves the technical problem of low inference efficiency of computational graphs.

(57) 摘要: 本申请公开了一种计算图的推理方法、系统及存储介质。其中, 该方法包括: 监测待推理的原始类型计算图; 响应于监测到的原始类型计算图转换为对应的目标类型计算图失败, 将原始类型计算图划分为一级子原始类型计算图; 将一级子原始类型计算图转化为对应的至少一目标类型计算图; 将与一级子原始类型计算图对应的目标类型计算图, 生成待推理文件, 且将待推理文件发送至远程设备上。本申请解决了计算图的推理效率低的技术问题。

[见续页]

MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。